2002s-50

Forecasting Non-Stationary Volatility with Hyper-Parameters

Yoshua Bengio, Charles Dugas

Série Scientifique
Scientific Series

CIRANO
Centre interuniversitaire de recherche en analyse des organisations

Montréal
Mai 2002

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de la Recherche, de la Science et de la Technologie, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère de la Recherche, de la Science et de la Technologie, and grants and research mandates obtained by its research teams.

Les organisations-partenaires / The Partner Organizations

- •École des Hautes Études Commerciales
- •École Polytechnique de Montréal
- •Université Concordia
- •Université de Montréal
- •Université du Québec à Montréal
- •Université Laval
- •Université McGill
- •Ministère des Finances du Québec
- •MRST
- •Alcan inc.
- •AXA Canada
- •Banque du Canada
- •Banque Laurentienne du Canada
- •Banque Nationale du Canada
- •Banque Royale du Canada
- •Bell Canada
- •Bombardier
- •Bourse de Montréal
- •Développement des ressources humaines Canada (DRHC)
- •Fédération des caisses Desjardins du Québec
- •Hydro-Québec
- •Industrie Canada
- •Pratt & Whitney Canada Inc.
- •Raymond Chabot Grant Thornton
- •Ville de Montréal

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.
This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Forecasting Non-Stationary Volatility with Hyper-Parameters*

Yoshua Bengio† and Charles Dugas‡

Résumé /Abstract

Nous considérons des données séquentielles échantillonnées à partir d'un processus inconnu, donc les données ne sont pas nécessairement iid. Nous développons une mesure de généralisation pour de telles données et nous considérons une approche récemment proposée pour optimiser les hyper-paramètres qui est basée sur le calcul du gradient d'un critère de sélection de modèle par rapport à ces hyper-paramètres. Les hyper-paramètres sont utilisés pour donner différents poids dans la séquence de données historiques. Notre approche est appliquée avec succès à la modélisation de la volatilité des rendements d'actions canadiennes sur un horizon de un mois.

We consider sequential data that is sampled from an unknown process, so that the data are not necessarily iid. We develop a measure of generalization for such data and we consider a recently proposed approach to optimizing hyper-parameters, based on the computation of the gradient of a model selection criterion with respect to hyper-parameters. Hyper-parameters are used to give varying weights in the historical data sequence. The approach is successfully applied to modeling the volatility of Canadian stock returns one month ahead.

Keywords: Sequential data, hyper-parameters, generalization, stock returns, volatility.

Mots-clés : Données séquentielles, hyper-paramètres, généralisation, rendement d'actions, volatilité.

* The authors would like to thank René Garcia, Réjean Ducharme, as well as the NSERC Canadian funding agency.

† CIRANO and Département d'informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Tel: +1 (514) 343-6804, email: bengioy@iro.umontreal.ca

‡ CIRANO and Département d'informatique et recherche opérationnelle, Université de Montréal, Montréal, Québec, Canada, H3C 3J7. Tel: +1 (514) 343-6804, email: bengioy@iro.umontreal.ca

1 Introduction

Many learning algorithms can be formulated as the minimization of a *training criterion* which involves both training errors on each training example and some *hyper-parameters*, which are kept fixed during this minimization. For example, in the regularization framework (Tikhonov & Arsenin, 1977), one hyper-parameter controls the strength of the penalty term and thus the capacity (Vapnik, 1998) of the system. Many criteria for choosing the value of the hyper-parameter have been proposed in the model selection literature, in general based on an estimate or a bound on generalization error (Vapnik, 1998; Akaike, 1974; Craven & Wahba, 1979). In (Bengio, 1999) we have introduced a new approach to simultaneously optimize many hyper-parameters, based on the computation of the *gradient of a model selection criterion with respect to the hyper-parameters*. In this paper, we apply this approach to the modeling of non-stationary time-series data, and present comparative experiments on financial returns data for Canadian stocks.

Most approaches to machine learning and statistical inference from data assume that data points are *i.i.d.* (see e.g. (Vapnik, 1998), but an exception is (Littlestone & Warmuth, 1994)). We will not consider bounds on generalization error, but simply estimates of generalization error and ways to optimize them explicitly. In particular, we use an extension of the cross-validation criterion that can be applied to sequential non-*i.i.d.* data. However, the general approach could be applied (and maybe better results obtained) with other model selection criteria, since we suspect that cross-validation estimates are very noisy. In Section 2, we formalize a notion of generalization error for data that are not *i.i.d.*, similar to the notion proposed in (Evans, Rajagopalan, & Vazirani, 1993), and describe an analogue to cross-validation to estimate this generalization error, that we call **sequential validation**. In Section 3, we summarize the theoretical results obtained in (Bengio, 1999) for computing the gradient of a model selection criterion with respect to hyper-parameters. In particular, we consider hyper-parameters that smoothly control what weight to give to past historical data. In Section 4, we describe experiments performed on artificial data that are generated by a non-stationary process with an abrupt change in distribution. In Section 5, we describe experiments performed on financial data: predicting next month's return first and second moment, for Canadian stocks. The results show that statistically significant improvements can be obtained with the proposed approach.

2 Generalization Error for Non-IID Data

Let us consider a sequence of data points Z_1, Z_2, ..., with $Z_t \in \mathcal{O}$ (e.g., $\mathcal{O} =$ the reals $\mathcal{R}^n$) generated by an unknown non-stationary process: $Z_t \sim P_t(Z)$. At each discrete time step t, in order to take a decision or make a prediction, we are allowed to choose a function f from a set of functions $\mathcal{F}$, and to do

this we are allowed to use past observations $D_t = (z_1, z_2, \dots z_t)$. Let us call this choice $f^*_{D_t}$. At the next time step (or more generally at some time step in the future), we will be able to evaluate the quality of our choice $f^*_{D_t}$, with a known cost function $Q(f^*_{D_t}, Z_{t+1})$. For example, we will describe experiments for the case of affine functions, i.e., $\mathcal{F} = \{f : \mathcal{R}^n \to \mathcal{R}^m | f(\mathbf{x}) = \theta\tilde{\mathbf{x}}, \mathbf{x} \in \mathcal{R}^n, \theta \in \mathcal{R}^{m\times(n+1)}, \tilde{\mathbf{x}} = (\mathbf{x}, 1) = (x_1, \dots, x_n, 1)\}$, and these functions will be selected to minimize the *squared loss*:

$$Q(f, (x, y)) = \frac{1}{2}(f(\mathbf{x}) - \mathbf{y})^2. \tag{1}$$

In this context, we can define the *expected generalization error* C_{t+1} at the next time step as the expectation over the unknown process P_{t+1} of the squared loss function Q:

$$C_{t+1}(f) = \int_{z_{t+1}} Q(f, z_{t+1})\, dP_{t+1}(z_{t+1}). \tag{2}$$

We would like to select f^* which has the lowest expected generalization error, but only an approximation of it can be reached. At time t, we are only given partial information on the process density P_t. We thus revert to the *empirical error* of the choice f, when data D_t have been observed which is

$$\widehat{C}_{t+1}(f, D_t) = \frac{1}{t}\sum_{s=1}^{t} Q(f, z_s). \tag{3}$$

The minimization of this empirical error would lead us to choose some "good" $f \in \mathcal{F}$. However, if $\mathcal{F}$ has too much capacity (Vapnik, 1998), it might be better to choose an f that minimizes an alternative functional, e.g., a *training criterion* that penalizes the complexity of f in order to avoid over-fitting the data and thus, avoid poor generalization:

$$\widehat{C}_{t+1}(f, D_t, \lambda) = R(f, \lambda) + \frac{1}{t}\sum_{s=1}^{t} Q(f, z_s) \tag{4}$$

where $\lambda \in \mathcal{R}^l$ is a vector of so-called **hyper-parameters** and $R(f, \lambda)$ is a penalty term that defines a preference over functions f within $\mathcal{F}$. A common heuristic used by practitioners of financial prediction is to train a model based only on recent historical data, because of the believed non-stationarity of this data. This heuristic corresponds to the assumption that there may be drastic changes in the distribution of financial and economic variables. More generally, the cost function can weight differently past observations, giving more weight to observations that occurred after the most recent change in the underlying process.

$$\widehat{C}_{t+1}(f, D_t, \lambda) = R(f, \lambda) + \frac{1}{t}\sum_{s=1}^{t} w_s(t, \lambda) Q(f, z_s) \tag{5}$$

where $w_s(t, \lambda)$ is a scalar function of λ that weights data observed at time s, using information up to time t. In this paper, we will concentrate on different weighting of past observations and drop the penalty term. Our functional is then:

$$\widehat{C}_{t+1}(f, D_t, \lambda) = \frac{1}{t} \sum_{s=1}^{t} w_s(t, \lambda) Q(f, z_s) \tag{6}$$

Let us suppose that one chooses $f^*_{D_t,\lambda}$ which minimizes the above $\widehat{C}_{t+1}$, given a fixed choice of λ and a set of data D_t. We obtain

$$f^*_{D_t,\lambda} = \arg\min_{f \in \mathcal{F}} \widehat{C}_{t+1}(f, D_t, \lambda) \tag{7}$$

One way to select the hyper-parameters λ in equations (6) and (7) is to consider *what would have been the generalization error* in the past if we had used the hyper-parameters λ. We will use a **sequential cross-validation** criterion:

$$\widehat{E}_{t+1}(\lambda, D_t) = \frac{1}{t - M'} \sum_{s=M'}^{t-1} Q(f^*_{D_s,\lambda}, z_{s+1}) \tag{8}$$

where $f^*_{D_s,\lambda}$ minimizes the training criterion (equation (6)), and M' is the minimum number of training points to "reasonably" select a value of f within $\mathcal{F}$. Our objective is to select the combination of parameters and hyper-parameters that minimizes the sequential cross-validation criterion. **At each time** t, we will select the hyper-parameters that minimize $\widehat{E}_{t+1}$:

$$\lambda^*_{D_t} = \arg\min_{\lambda} \widehat{E}_{t+1}(\lambda, D_t) \tag{9}$$

This is similar to the principle of minimization of the empirical error, but applied to the selection of hyper-parameters, in the context of sequential data. Once $\lambda^*_{D_t}$ has been selected, the corresponding $f^*_{D_t}$ is therefore obtained:

$$f^*_{D_t} = f^*_{D_t,\lambda^*_{D_t}} = \arg\min_{f \in \mathcal{F}} \widehat{C}_{t+1}(f, D_t, \lambda^*_{D_t}). \tag{10}$$

Let M be the "minimum" number of training points for learning both parameters and hyper-parameters. Let T be the total number of data points. We estimate the generalization error of the system that chooses the set $\{f^*_{D_s}\}$, $s \in \{M', M'+1, \ldots T\}$ as follows:

$$\widehat{G}_{t+1}(D_T) = \frac{1}{T - M} \sum_{s=M}^{T-1} Q(f^*_{D_s}, z_{s+1}) \tag{11}$$

One might wonder why the generalization error estimate is computed using the set of functions $\{f^*_{D_s}\}$ instead of the very last estimated function $f^*_{D_T}$: the choice of hyper-parameters at t should only depend on data available up to time t. The generalization error must therefore be seen as a generalization error for a "strategy", rather than the generalization error of a single, particular function.

3 Optimizing Hyper-Parameters for Non-IID Data

In this section we summarize the theoretical results already presented in (Bengio, 1999) and extend them for the application of interest here, i.e., using hyper-parameters for modeling possibly non-stationary time-series.

If the functions of $\mathcal{F}$ are smooth in their parameters, the optimized θ depends on the choice of hyper-parameters in a continuous way:

$$\theta^*(D_t, \lambda) = \arg\min_{\theta} \widehat{C}_{t+1}(\theta, D_t, \lambda). \tag{12}$$

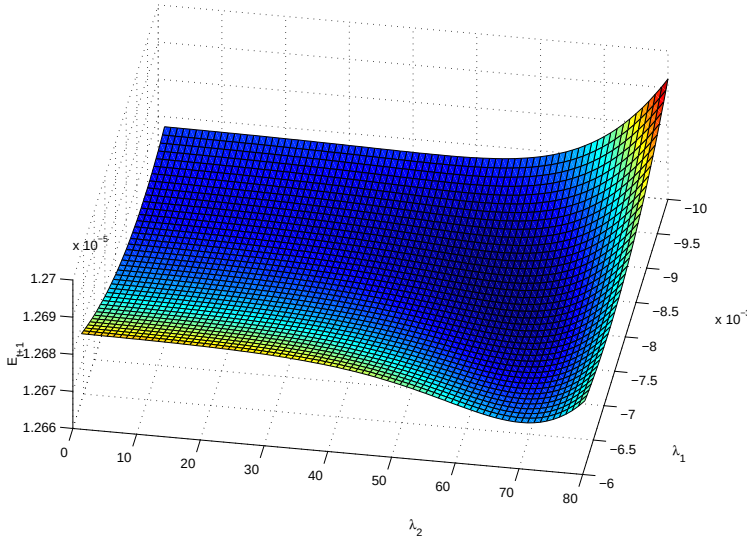

Figure 1: Values of the sequential selection criterion for various values of the hyper-parameters. As in this particular case, the function is generally smooth. Here, the global minimum is obtained with $\lambda_1 = -0.0085$ and $\lambda_2 = 60$.

Note that there is a bijection between a parameterized function f and its set of parameters θ. Therefore, the above equation (12) is the same as the previous equation (7). For the application considered in this paper, the hyper-parameters λ are used for controlling *training weights* on past data points, i.e., giving a weight $w_s(t, \lambda)$ to the past observation at time s when minimizing the training error up to time t, in equation (6). To weight the past data points, we have considered as an example a sigmoidal decay, with hyper-parameters λ_1 and λ_2, (see the smooth curve of weights in Figure 2):

$$w_s(t,\lambda) = \frac{1}{1+\exp(-\lambda_1(s-\lambda_2))} \tag{13}$$

This is a smooth version of the "abrupt transition" heuristic used in practice.

To apply gradient-based optimization to the selection of the hyper-parameters (equation (9)), we will compute the gradient of the sequential cross-validation criterion $\widehat{E}_{t+1}$ (equation (8)) with respect to the hyper-parameters:

$$\left.\frac{\partial \widehat{E}_{t+1}}{\partial \lambda}\right|_{\lambda=\lambda_0} = \frac{1}{t-M'} \cdot \sum_{s=M'}^{t-1} \left.\frac{\partial Q(\theta, z_{s+1})}{\partial \theta}\right|_{\theta=\theta^*(D_s,\lambda_0)} \cdot \left.\frac{\partial \theta^*(D_s,\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} \tag{14}$$

Basically, this involves looking at the minimum of $\widehat{C}_{t+1}$ with respect to the parameter vector θ, for a given hyper-parameter vector λ (equation (12)), and then seeing how a change in λ influences the solution θ. The latter is rather unusual as we need to compute the derivative of the parameters θ with respect to the hyper-parameters λ.

Let us first differentiate the value of the derivative of the quadratic cost function with respect to the parameters. If $f \in \mathcal{F}$ is affine, then Q is simply a quadratic function of the parameters and can thus be rewritten as

$$Q = a(\lambda) + b(\lambda)\theta + \frac{1}{2}\theta' H(\lambda)\theta \tag{15}$$

where λ is considered fixed and $H(\lambda)$ is the symmetric positive-definite Hessian matrix of second derivatives of the training criterion with respect to the parameters. The computation of gradients with respect to hyper-parameters can also be performed for non-quadratic cost functions (Bengio, 1999) but in this paper, we only need to consider the simpler quadratic case. Differentiating Q with respect to θ, we obtain:

$$\frac{\partial Q}{\partial \theta} = b(\lambda) + H(\lambda)\theta \tag{16}$$

We now need to obtain the values of $b(\lambda)$ and $H(\lambda)$. Since,

$$b(\lambda) = \left.\frac{\partial Q}{\partial \theta}\right|_{\theta=0}, \tag{17}$$

we estimate $b(\lambda)$ and $H(\lambda)$ as:

$$b_s(\lambda_0) = -\frac{1}{s}\sum_{u=1}^{s} w_u(s,\lambda_0)\cdot \tilde{x}_u \cdot y'_u \tag{18}$$

$$H_s(\lambda_0) = \frac{1}{s}\sum_{u=1}^{s} w_u(s,\lambda_0)\cdot \tilde{x}_u \cdot \tilde{x}'_u \tag{19}$$

Setting equation (16) to zero, we obtain

$$\theta^*(D_s, \lambda_0) = -H_s^{-1}(\lambda_0) \cdot b_s(\lambda_0) \tag{20}$$

And so, we are now able to compute the first term in equation (14):

$$\left.\frac{\partial Q(\theta, z_{s+1})}{\partial \theta}\right|_{\theta=\theta^*(D_s,\lambda_0)} = (\theta^*(D_s, \lambda_0) \cdot \tilde{x}_{s+1} - y_{s+1}) \cdot \tilde{x}_{s+1} \tag{21}$$

Let us now derive the value of the derivative of the optimal parameters with respect to the hyper-parameters. Equation (20) provides us with an expression of the optimal parameters as a function of the hyper-parameters which we only need to differentiate to obtain the solution for the second term in the summation in equation (14):

$$\left.\frac{\partial \theta^*(D_s, \lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} = -\left.\frac{\partial H^{-1}(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} \cdot b_s(\lambda_0) - \left.\frac{\partial b(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} \cdot H_s^{-1}(\lambda_0) \tag{22}$$

The values of $H_s^{-1}(\lambda_0)$ and $b_s(\lambda_0)$ are given by equations (18) and (19). Computing the derivative of these functions with respect to the hyper-parameters gives us the value of the two remaining terms.

$$\left.\frac{\partial b(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} = -\frac{1}{s}\sum_{u=1}^{s} \left.\frac{\partial w_u(s, \lambda)}{\partial \lambda}\right|_{\lambda_0} \cdot \tilde{x}_u \cdot y'_u \tag{23}$$

$$\left.\frac{\partial H(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} = \frac{1}{s}\sum_{u=1}^{s} \left.\frac{\partial w_u(s, \lambda)}{\partial \lambda}\right|_{\lambda_0} \cdot \tilde{x}_u \cdot \tilde{x}'_u \tag{24}$$

Noting that

$$\frac{\partial H^{-1}}{\partial \lambda} H + \frac{\partial H}{\partial \lambda} H^{-1} = \frac{\partial H^{-1} H}{\partial \lambda} = \frac{\partial I}{\partial \lambda} = 0 \tag{25}$$

We can isolate $\frac{\partial H^{-1}}{\partial \lambda}$ as a function of known values:

$$\frac{\partial H^{-1}}{\partial \lambda} = -H^{-1} \frac{\partial H}{\partial \lambda} H^{-1} \tag{26}$$

An even more efficient method using the Cholesky decomposition of the Hessian matrix can be found in (Bengio, 1999).

All four terms of equation (22) are known. We can then use the results of equations (22) and (21) to compute the derivative of the sequential cross-validation criterion with respect to the hyper-parameters (equation (14)). Then, using gradient descent will allow us to search the space of hyper-parameters in a continuous fashion towards a minimum (possibly local) of the sequential cross-validation criterion. Then, equation (11) provides us with an estimate of the generalization error of the whole process.

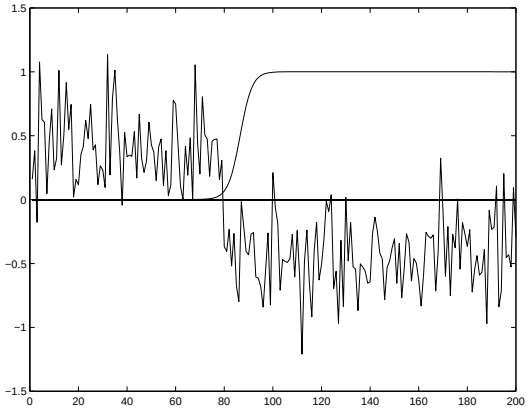

Figure 2: Ratio $\frac{y_t}{x_t}$ vs t from the generated data (noisy curve), and training weights $w_s(200, \lambda)$ (smooth curve) found at the end of the sequence by optimizing the hyper-parameters λ_1 and λ_2.

4 Experiments on Artificial Data

We have tested the algorithm on artificially generated non-*i.i.d.* data with a single abrupt change in the input/output dependency at some point in the sequence. The single input/single output data sequence is generated by a Gaussian mixture for the inputs, and for the output given the input a left-to-right Input/Output Hidden Markov Model (Bengio & Frasconi, 1996). A sequence of $T = 200$ input/output pairs was generated. We set $M' = 25$ and $M = 50$ (see equations (8) and (11)).

At time $t = 80$, when the model randomly switches to a state with a different distribution, the relation between the input and the output changes drastically, as seen in Figure 2 (with the ratio of output to input noisy curve).

The smooth curve in Figure 2 illustrates the values of the pattern weights after the 200 observations have been taken account of. Clearly, the algorithm has succeeded in identifying the input/output dependency change at time $t = 80$. Accordingly, those observations after time $t = 80$ are given weights close to 1 and observations before the transition point are given close to null weighting. Whereas research often concentrates on selecting, among a group of candidates, an optimal period of time which is to be used for all time-series of a given type and for the purpose of prediction, our algorithm selects this period automatically and independently for each time-series, thus recognizing that non-stationarities are unlikely to occur strictly simultaneously for all processes.

We compare a regular linear regression with a linear regression with three hyper-parameters: one hyper-parameter for the weight decay and two hyper-parameters (λ_1 and λ_2) to yield training weights on past data points with a sigmoidal decay (equation (13)). The out-of-sample MSE for the regular linear regression is 3.97 whereas it is only 0.62 when the hyper-parameters are opti-

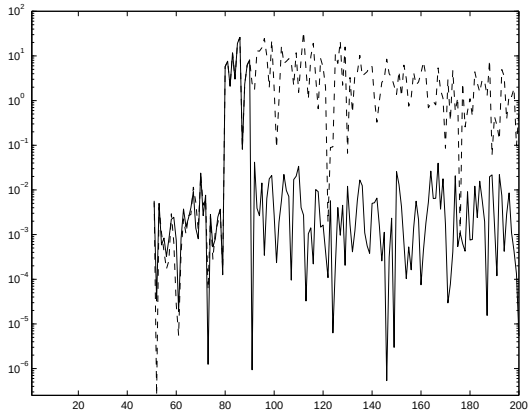

Figure 3: Out-of-sample squared loss for each time step t, for the ordinary regression (dashed) and regression with optimized hyper-parameters (continuous).

mized. As shown in Figure 3, much better performance is obtained with the adaptive hyper-parameters, which allow to **quickly recover from the change in distribution** at $t = 80$. Between times $t = 0$ and $t = 80$, both errors closely map one another at low values. Then from time $t = 80$ to time $t \approx 90$, both errors are are much higher (note the log-scale plot). Afterward, the model with hyper-parameters recovers and its errors drop back to values close to the ones obtained before the abrupt change. On the other side, the plain vanilla linear regression, stays stocked at high error values, unable to detect the change.

5 Experiments on Financial Time-Series

In this section, we describe experiments performed on financial data: predicting next month's return first and second moment, for Canadian stocks. Let

$$r_t = value_t / value_{t-1} - 1$$

be a discrete return series (the ratio of the value of an asset at time t over its value at time $t-1$, which in the case of stocks includes dividends and capital gains). *In these experiments, our goal is to make predictions on the first and second moment of r_{t+1}, using information available at time t.* These predictions could be used in financial decision taking in various ways: for asset allocation (taking risks into account), for estimating risks, and for pricing derivatives (such as options, whose price depends on the second moment of the returns).

In the experiments we directly train our models to predict these two moments by minimizing the squared error, i.e., we are trying to learn $E_t[r_{t+1}]$ and $E_t[r_{t+1}^2]$ where E_t denotes the conditional expectation using information available up to time t.

We have performed experiments on monthly returns and monthly squared returns of 473 stocks from the Toronto Stock Exchange (TSE) for which at least 98 months of data were available, the earliest starting in January 1976, and the latest data ending in December 1996.

5.1 Experimental Setup and Performance Measures

In all the experiments, we compare several models on the same data, and we use the *sequential cross-validation* criterion (equation (11)) to estimate generalization performance. For models with hyper-parameters, the "minimum" number of training points to evaluate parameters while hyper-parameters remained fixed was set to $M' = 48$ months (4 years). The minimum number of training points in order to train hyper-parameters was set to $M = 72$ months (6 years). In all other experiments using models without hyper-parameters, the "minimum" number of training points was set to $M = 72$. The out-of-sample MSE was computed for each stock. The results reported below concern the average MSE over all the 473 stocks. We have also estimated the variance across stocks of the MSE value, and the variance across stocks of the difference between the MSE for one model and the MSE for a reference model, as described below. Using the latter, we have tested the null hypothesis that two compared models have identical true generalization error.

For this purpose, we have used an estimate of variance that takes into account the autocorrelation of errors through time. Let e_t be a series of errors (e.g. squared prediction error) with sample mean $\bar{e}$.

$$\bar{e} = \frac{1}{n} \sum_{t=1}^{n} e_t \tag{27}$$

We are interested in estimating its variance:

$$Var[\bar{e}] = \frac{1}{n^2} \sum_{t=1}^{n} \sum_{t'=1}^{n} Cov(e_t, e_{t'}). \tag{28}$$

Since we are dealing with a time-series and because we do not know how to estimate independently and reliably all the above covariances, we will assume that the error series is covariance-stationary and that the covariance dies out as $|t-t'|$ increases. This can be verified empirically by drawing the autocorrelation function of the e_t series. The covariance stationarity implies that

$$Cov(e_t, e_{t'}) = \gamma_{|t-t'|}, \tag{29}$$

where the γ's are estimated from the sample covariances:

$$\hat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (e_t - \bar{e})(e_{t+k} - \bar{e}) \tag{30}$$

An unbiased and convergent estimator of this variance is thus the following (Priestley, 1981):

$$\widehat{Var}[\bar{e}] = \frac{1}{n}\left(\hat{\gamma}_0 + 2 \cdot \sum_{k=1}^{m}(1 - k/m)\hat{\gamma}_k\right), \tag{31}$$

where $\lim_{n\to\infty} m = \infty$ and $\lim_{n\to\infty} m/n = 0$: we have used $m = \sqrt{n}$.

Because there are generally strong dependencies between the errors of different models, we have found that much better estimates of variance were obtained by analyzing the differences of squared errors, rather than computing a variance separately for each average:

$$Var[\bar{e}^A - \bar{e}^B] = Var[\bar{e}^A] + Var[\bar{e}^B] - 2Cov[\bar{e}^A, \bar{e}^B] \tag{32}$$

In the tables below we give the p-value of the null hypothesis that a model is not better than the reference model, using a normal approximation for the differences in average sequence errors.

5.2 Models Compared in the Experiments

The following models have been considered in the comparative experiments (note the "short name", in bold below, used in the tables):

- **Constant** model: this is the *reference* or *naive* model, which has *1 free parameter*, the *historical average* of the past and current output observations.
- **Linear 1** model is a linear regression with 1 input, which is the current value y_t of the output variable y_{t+1}; it has *2 free parameters*.
- **Linear 2** model is **Linear 1** model with an extra input which is the average over the last 6 months of the output variable; it has *3 free parameters*.
- **Linear 4** model is **Linear 2** model, with two extra inputs: the value of the output variable at the previous time step y_{t-1} and the average over the last 6 time steps of the return; this model was only used for the squared return prediction experiments; it has *5 free parameters*.
- **ARMA(p,q)** model is a recurrent model of orders p (auto-regressive recurrences) and q (moving average lags). It has *1+p+q free parameters*; we have tried the following combinations of p and q: (1,1),(2,1),(1,2),(2,2).
- **Hyper Constant** model is the **Constant** model with weights on the past data (learned with the hyper-parameters); there is *1 free parameter* and *2 hyper-parameters*.
- **Hyper Linear 1** model is the **Linear 1** model with weights on the past data; there are *2 free parameters* and *2 hyper-parameters*.

5.3 Experimental Results

The results on predicting the first moment of next month's stocks returns are given in table 1. Note that the p-values are one-sided but that we do two different tests depending on whether the tested model average error is less or greater than the reference (Constant model). In the latter case there is a "W" in the p-value column, indicating that the performance is worse than the reference. Significant results at the 5% level are indicated with a bold p-value. The lowest error over all the models is indicated by a bold MSE. For the first moment, the constant model significantly beats all the others. The results on predicting the second moment are given in table 2. The constant model with hyper-parameters significantly beats all the others.

	MSE	(sdev)	p-value
Constant	**7.779e-3**	(2.02e-4)	
Hyper Constant	8.075e-3	(2.13e-4)	**W <1e-7**
Hyper Linear 1	9.484e-3	(5.20e-4)	**W1.35e-5**
Linear 1	7.884e-3	(2.08e-4)	**W <1e-7**
ARMA(1,1)	7.908e-3	(2.12e-4)	**W <1e-7**

Table 1: Results of experiments on predicting one-month ahead **stocks returns** using a variety of models. The average out-of-sample squared error times 0.5 (MSE) over all the assets are given, with estimated standard deviation of the average in parentheses, and p-value of the null hypothesis of no difference with the Constant model. A "W" means that the alternative hypothesis is that the model is WORSE than the constant model. *All the models are significantly worse than the **Constant** model.*

6 Conclusions

In this paper we have achieved the following: (1) We have introduced an extension of cross-validation, called **sequential cross-validation** as a model selection criterion for possibly non-*i.i.d.* data. (2) We have applied the method for optimizing hyper-parameters introduced in (Bengio, 1999) to the special case of capturing abrupt changes in non-stationary data: 2 hyper-parameters control the weight on each past time step. (3) We have tested the method on artificial data, showing that when there is such an abrupt change, the regression is much improved by using and optimizing the hyper-parameters. (4) We have tested the method on financial returns data to predict the first and second moment of next month's return for individual stocks. The specific conclusions of these experiments are the following: On estimating the conditional expectation of stock returns, the constant model significantly beats all the tested models, including linear and ARMA models. On estimating the conditional expectation of the squared return (which can be used to predict volatility), the constant mod-

	MSE	(sdev)	p-value
Constant	3.282e-3	(2.02e-4)	
Hyper Constant	**3.215e-3**	(2.13e-4)	**0.0105**
Linear 1	5.208e-3	(1.87e-3)	W0.117
Linear 2	5.707e-3	(1.82e-3)	W0.0522
Linear 4	5.342e-3	(1.89e-3)	W0.0994
ARMA(1,1)	5.617e-3	(2.22e-3)	**W1.94e-3**
ARMA(2,1)	5.515e-3	(2.08e-3)	**W1.79e-3**
ARMA(1,2)	5.910e-3	(2.30e-3)	**W2.01e-3**
ARMA(2,2)	5.637e-3	(1.97e-3)	**W1.66e-3**

Table 2: Results of experiments on predicting one-month ahead **stocks squared returns** using a variety of models. The average out-of-sample squared error times 0.5 (MSE) over all the assets are given, with estimated standard deviation of the average in parentheses, and p-value of the null hypothesis of no difference with the Constant model. A "W" means that the alternative hypothesis is that the model is WORSE than the constant model. *The **Hyper Constant** model is significantly better than all the others while the **ARMA** models are all significantly worse than the Constant reference.*

el with hyper-parameters to handle non-stationarities beats the other models, with a p-value of 1%.

What remains to be done, in the direction of research that we have explored here? In our experiments we have found that the estimates of the hyper-parameters was very sensitive to the data, probably because of the variance of the cross-validation criterion, so better results might be obtained by using a less noisy model selection criterion. It would also be interesting to see if the significant improvements that we have found for Canadian stocks can be observed on other markets, and if these predictions could be used to improve specific decisions concerning those stocks (such as for trading options).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*(6), 716–728.

Bengio, Y. (1999). Continuous optimization of hyper-parameters. Tech. rep. 1144, Département d'informatique et recherche opérationnelle, Université de Montréal.

Bengio, Y., & Frasconi, P. (1996). Input/Output HMMs for sequence processing. *IEEE Transactions on Neural Networks*, *7*(5), 1231–1249.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematics*, *31*, 377–403.

Evans, W., Rajagopalan, S., & Vazirani, U. (1993). Choosing a reliable hypothesis. In *Proceedings of the 6th Annual Conference on Computational Learning Theory*, pp. 269–276 Santa Cruz, CA, USA. ACM Press.

Littlestone, N., & Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, *108*(2), 212–261.

Priestley, M. (1981). *Spectral Analysis and Time Series, Vol.1: Univariate Series*. Academic Press.

Tikhonov, A., & Arsenin, V. (1977). *Solutions of Ill-posed Problems*. W.H. Winston, Washington D.C.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Liste des publications au CIRANO*

Série Scientifique / *Scientific Series* (ISSN 1198-8177)

2002s-50 Forecasting Non-Stationary Volatility with Hyper-Parameters / Y. Bengio et C. Dugas

2002s-49 Cost Functions and Model Combination for VaR-based Asset Allocation using Neural Networks / N. Chapados et Y. Bengio

2002s-48 Experiments on the Application of IOHMMs to Model Financial Returns Series / Y. Bengio, V.-P. Lauzon et R. Ducharme

2002s-47 Valorisation d'Options par Optimisation du Sharpe Ratio / Y. Bengio, R. Ducharme, O. Bardou et N. Chapados

2002s-46 Incorporating Second-Order Functional Knowledge for Better Option Pricing / C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau et R. Garcia

2002s-45 Étude du biais dans le Prix des Options / C. Dugas et Y. Bengio

2002s-44 Régularisation du Prix des Options : Stacking / O. Bardou et Y. Bengio

2002s-43 Monotonicity and Bounds for Cost Shares under the Path Serial Rule / Michel Truchon et Cyril Téjédo

2002s-42 Maximal Decompositions of Cost Games into Specific and Joint Costs / Michel Moreaux et Michel Truchon

2002s-41 Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models / Sílvia Gonçalves, Halbert White

2002s-40 Selective Penalization Of Polluters: An Inf-Convolution Approach / Ngo Van Long et Antoine Soubeyran

2002s-39 On the Mediational Role of Feelings of Self-Determination in the Workplace: Further Evidence and Generalization / Marc R. Blais et Nathalie M. Brière

2002s-38 The Interaction Between Global Task Motivation and the Motivational Function of Events on Self-Regulation: Is Sauce for the Goose, Sauce for the Gander? / Marc R. Blais et Ursula Hess

2002s-37 Static Versus Dynamic Structural Models of Depression: The Case of the CES-D / Andrea S. Riddle, Marc R. Blais et Ursula Hess

2002s-36 A Multi-Group Investigation of the CES-D's Measurement Structure Across Adolescents, Young Adults and Middle-Aged Adults / Andrea S. Riddle, Marc R. Blais et Ursula Hess

2002s-35 Comparative Advantage, Learning, and Sectoral Wage Determination / Robert Gibbons, Lawrence F. Katz, Thomas Lemieux et Daniel Parent

2002s-34 European Economic Integration and the Labour Compact, 1850-1913 / Michael Huberman et Wayne Lewchuk

2002s-33 Which Volatility Model for Option Valuation? / Peter Christoffersen et Kris Jacobs

* Consultez la liste complète des publications du CIRANO et les publications elles-mêmes sur notre site Internet :

http://www.cirano.qc.ca

2002s-32 Production Technology, Information Technology, and Vertical Integration under Asymmetric Information / Gamal Atallah

2002s-31 Dynamique Motivationnelle de l'Épuisement et du Bien-être chez des Enseignants Africains / Manon Levesque, Marc R. Blais, Ursula Hess

2002s-30 Motivation, Comportements Organisationnels Discrétionnaires et Bien-être en Milieu Africain : Quand le Devoir Oblige / Manon Levesque, Marc R. Blais et Ursula Hess

2002s-29 Tax Incentives and Fertility in Canada: Permanent vs. Transitory Effects / Daniel Parent et Ling Wang

2002s-28 The Causal Effect of High School Employment on Educational Attainment in Canada / Daniel Parent

2002s-27 Employer-Supported Training in Canada and Its Impact on Mobility and Wages / Daniel Parent

2002s-26 Restructuring and Economic Performance: The Experience of the Tunisian Economy / Sofiane Ghali and Pierre Mohnen

2002s-25 What Type of Enterprise Forges Close Links With Universities and Government Labs? Evidence From CIS 2 / Pierre Mohnen et Cathy Hoareau

2002s-24 Environmental Performance of Canadian Pulp and Paper Plants : Why Some Do Well and Others Do Not ? / Julie Doonan, Paul Lanoie et Benoit Laplante

2002s-23 A Rule-driven Approach for Defining the Behavior of Negotiating Software Agents / Morad Benyoucef, Hakim Alj, Kim Levy et Rudolf K. Keller

2002s-22 Occupational Gender Segregation and Women's Wages in Canada: An Historical Perspective / Nicole M. Fortin et Michael Huberman

2002s-21 Information Content of Volatility Forecasts at Medium-term Horizons / John W. Galbraith et Turgut Kisinbay

2002s-20 Earnings Dispersion, Risk Aversion and Education / Christian Belzil et Jörgen Hansen

2002s-19 Unobserved Ability and the Return to Schooling / Christian Belzil et Jörgen Hansen

2002s-18 Auditing Policies and Information Systems in Principal-Agent Analysis / Marie-Cécile Fagart et Bernard Sinclair-Desgagné

2002s-17 The Choice of Instruments for Environmental Policy: Liability or Regulation? / Marcel Boyer, Donatella Porrini

2002s-16 Asymmetric Information and Product Differentiation / Marcel Boyer, Philippe Mahenc et Michel Moreaux

2002s-15 Entry Preventing Locations Under Incomplete Information / Marcel Boyer, Philippe Mahenc et Michel Moreaux

2002s-14 On the Relationship Between Financial Status and Investment in Technological Flexibility / Marcel Boyer, Armel Jacques et Michel Moreaux

2002s-13 Modeling the Choice Between Regulation and Liability in Terms of Social Welfare / Marcel Boyer et Donatella Porrini

2002s-12 Observation, Flexibilité et Structures Technologiques des Industries / Marcel Boyer, Armel Jacques et Michel Moreaux

2002s-11 Idiosyncratic Consumption Risk and the Cross-Section of Asset Returns / Kris Jacobs et Kevin Q. Wang

2002s-10 The Demand for the Arts / Louis Lévy-Garboua et Claude Montmarquette

2002s-09 Relative Wealth, Status Seeking, and Catching Up / Ngo Van Long, Koji Shimomura

2002s-08 The Rate of Risk Aversion May Be Lower Than You Think / Kris Jacobs

2002s-07 A Structural Analysis of the Correlated Random Coefficient Wage Regression Model / Christian Belzil et Jörgen Hansen

2002s-06 Information Asymmetry, Insurance, and the Decision to Hospitalize / Åke Blomqvist et Pierre Thomas Léger

2002s-05 Coping with Stressful Decisions: Individual Differences, Appraisals and Choice / Ann-Renée Blais

2002s-04 A New Proof Of The Maximum Principle / Ngo Van Long et Koji Shimomura

2002s-03 Macro Surprises And Short-Term Behaviour In Bond Futures / Eugene Durenard et David Veredas

2002s-02 Financial Asset Returns, Market Timing, and Volatility Dynamics / Peter F. Christoffersen et Francis X. Diebold

2002s-01 An Empirical Analysis of Water Supply Contracts / Serge Garcia et Alban Thomas

2001s-71 A Theoretical Comparison Between Integrated and Realized Volatilities Modeling / Nour Meddahi

2001s-70 An Eigenfunction Approach for Volatility Modeling / Nour Meddahi

2001s-69 Dynamic Prevention in Short Term Insurance Contracts / M. Martin Boyer et Karine Gobert

2001s-68 Serial Cost Sharing in Multidimensional Contexts / Cyril Téjédo et Michel Truchon

2001s-67 Learning from Strike / Fabienne Tournadre et Marie-Claire Villeval

2001s-66 Incentives in Common Agency / Bernard Sinclair-Desgagné

2001s-65 Detecting Mutiple Breaks in Financial Market Volatility Dynamics / Elena Andreou et Eric Ghysels

2001s-64 Real Options, Preemption, and the Dynamics of Industry Investments / Marcel Boyer, Pierre Lasserre, Thomas Mariotti et Michel Moreaux

2001s-63 Dropout, School Performance and Working while in School: An Econometric Model with Heterogeneous Groups / Marcel Dagenais, Claude Montmarquette et Nathalie Viennot-Briot

2001s-62 Derivatives Do Affect Mutual Funds Returns : How and When? / Charles Cao, Eric Ghysels et Frank Hatheway

2001s-61 Conditional Quantiles of Volatility in Equity Index and Foreign Exchange Data / John W. Galbraith, Serguei Zernov and Victoria Zinde-Walsh

2001s-60 The Public-Private Sector Risk-Sharing in the French Insurance "Cat. Nat. System" / Nathalie de Marcellis-Warin et Erwann Michel-Kerjan

2001s-59 Compensation and Auditing with Correlated Information / M. Martin Boyer et Patrick González

2001s-58 Resistance is Futile: An Essay in Crime and Commitment / M. Martin Boyer

2001s-57 The Unreliability of Output Gap Estimates in Real Time / Athanasios Orphanides et Simon van Norden